International Journal of Information Technology, Vol. 11, No. 7, 2005

The Confidence Measures for Isolated Word Recognition System

Yuguo Ding, Jia Liu, and Runsheng Liu

Department of Electronic Engineering, Tsinghua University
100084, Beijing, China
ding_ding98@mails.tsinghua.edu.cn
liuj@tsinghua.edu.cn
lrs-dee@mail.tsinghua.edu.cn

Abstract

This paper relates to the confidence measures for isolated word recognition system. Two types of confidence are introduced: online garbage model based likelihood ratio and semi-syllable based posterior probability. Linear classification is adopted to combine the two confidence scores. Traditional evaluation of confidence measure is adopted in the experiments. The experimental result shows, after the combining in the back-end processing, an acceptable performance is achieved for practical applications.

Keyword: Speech recognition, confidence measure, likelihood ratio, posterior probability.

1 Introduction

In decades, great improvements have been achieved in speech recognition. The technology has been already applied to practical systems[1]. The auto speech recognition (ASR) system has an outstanding performance for in-vocabulary (IV) input in office environments. But in practical system, it is important to notify a user of a rejection of an out-of-vocabulary (OOV) input or an unreliable result. The research and application of confident measure therefore is a critical aspect for practical speech recognition. Various methods have been proposed for computing confidence scores, include simple acoustic measures[2], N-best lists information[3] and combined measures[4].

In our system, we present two kinds of confidence measures: online garbage model based likelihood ratio and semi-syllable based posterior probability. In the back-end of our system, we combine two confidence measures according to Fisher Rule. The experimental results show a better performance after combination.

The paper is organized as follows: Section 2 describes two confidence measures in detail. The combination of confidence measures is introduced in section 3 and section 4 briefly gives the traditional evaluations of confidence measures. Section 5 is the experimental results and the conclusions are drawn in section 6.

2 Confidence Measures

In most of the current Speech Recognition System, the output result W is obtained according to the following maximum a posterior (MAP) decoder (1):

$$W^* = \arg\max_{W\in\Omega} P(W|X) \tag{1}$$

where Ω is the recognition set and X is the observed vectors.

According to Bayesian Formula, we have:

$$W^* = \arg\max_{W\in\Omega} \frac{P(X|W)P(W)}{P(X)} \tag{2}$$

While in a certain synchronous decoding, $P(X)$ is the same for all the lemmas of the recognition set and is always ignored. For the isolated word ASR system, $P(W)$ is also always ignored. So the base line of an ASR decoder is always described as follows:

$$W^* = \arg\max_{W\in\Omega} P(X|W) \tag{3}$$

From above we know, $P(X|W)$, which is always called likelihood, has no real means to evaluate accuracy of the best hypothesis.

2.1 Online Garbage Model Based Likelihood Ratio

In order to remedy the raw of the above decoding method, during the back-end processing, some other information is considered to generate likelihood ratio:

$$\frac{P(X \mid W_0)}{P(X \mid H_{Filler})} \tag{4}$$

where W_0 is the best hypothesis and H_{Filler} is called Filler model.

In our system, N hypotheses, $\{W_0, W_1, \cdots, W_{N-1}\}$, are listed out. Except the best hypothesis W_0, the excess hypotheses are called online garbages. The online garbages can represent the Filler model. Statistically, if the input speech data is in vocabulary, the likelihood of W_0 is distinctly larger than the online garbages. So, the denominator likelihood can be rewritten as follows:

$$P(X \mid H_{Filler}) = \sum_{i=1}^{N-1} P(X \mid W_i)P(W_i) \tag{5}$$

In the log domain, the normalized likelihood is as (6).

$$C_{LR}(W_0, X) = \frac{1}{n_x}[\log(P(X \mid W_0)) - \frac{1}{N-1}\sum_{i=1}^{N} \log(P(X \mid W_i))] \tag{6}$$

where n_x is the length of input speech in time domain. The likelihood ratio can describe how goodness of the match between input speech and the best hypothesis. And it is called online garbage model based confidence. The Online garbage model based likelihood ratio needs low additional computation and can
be simply realized.

2.2 Semi-syllable Based Posterior Probability Confidence Scores [5]

If we focus on the posterior probability, we can get another efficient confidence:

$$\begin{aligned} C_{PP}(W_0, X) &= P(W_0 \mid X) \\ &= \frac{P(X \mid W_0)}{P(X)} P(W_0) \end{aligned} \tag{7}$$

Considering the semi-syllable alignments of the observed vectors and assuming W_0 can be separated into M Semi-syllables $h_1, h_2, \cdots, h_M$, we can write the conditional probability $P(X \mid W_0)$ as:

$$P(X \left| W_0 \right.) = P(X_1, X_2, \cdots, X_M \left| h_1, h_2, \cdots, h_M \right.) \tag{8}$$

where h_i is the semi-syllable alignment of the observed vector sequence X_i, while the observe vector X is segmented during Viterbi match. Assuming the observed vector X_i is independent of each other. We have:

$$P(X) = \prod_{\mathrm{i=1}}^{\mathrm{m}} \mathrm{P}(X_i) \tag{9}$$

If we choose a unigram language model, we can rewrite the conditional probability $P(X \mid W_0)$ and the language probability $P(W_0)$ as follows:

$$\begin{aligned} P(X \left| W_0 \right.) &= P(X_1, X_2, \cdots, X_M \left| h_1, h_2, \cdots, h_M \right.) \\ &= \prod_{\mathrm{i=1}}^{\mathrm{m}} \mathrm{P}(X_i \left| h_i \right.) \end{aligned} \tag{10}$$

$$\begin{aligned} P(W_0) &= P(h_1, h_2, \cdots, h_M) \\ &= \prod_{\mathrm{i=1}}^{\mathrm{m}} \mathrm{P}(h_i) \end{aligned} \tag{11}$$

So, we get

$$C_{PP}(W_0, X) = \prod_{i=1}^{m} \frac{P(X_i \mid h_i) P(h_i)}{P(X_i)} \tag{12}$$

For each segmented observed vector X_i,

$$P(X_i) = \sum_j P(X_i \mid h_j) P(h_j) \tag{13}$$

Consequently, (7) can be described as:

$$\begin{aligned} C_{PP}(W_0, X) &= P(W_0 \mid X) \\ &= \prod_{i=1}^{m} \frac{P(X_i \mid h_i) P(h_i)}{\sum_j P(X_i \mid h_j) P(h_j)} \end{aligned} \tag{14}$$

In our system, we choose monophone as the acoustic model and unigram as the language model. For Chinese mandarin, each word consists of consonant and vowel. Consequently, h_i is matched as consonant, $\left\{h_j\right\}$ in the denominator involves all the consonants. Otherwise $\left\{h_j\right\}$ involves all the vowels.

3 Combinations of Two Confidence Scores

Since the two kinds of confidence scores are of different information, even better performance will be achieved if combining them together. Neural network, SVM and Linear classification can be considered. For computational reasons, we use Fisher linear classification to combine the two confidence scores:

$$CM(W_0, X) = \alpha C_{LR}(W_0, X) + \beta C_{PP}(W_0, X) \tag{15}$$

The coefficients α and β in (15) are determined according to Fisher Rule[6]. Sufficient data should be prepared to generate the training scores of the IV and OOV utterances. For N_1 times IV input, we get

$$x_i = \left[C_{LR} \quad C_{PP} \right], i = 1, 2, \cdots, N_1 \tag{16}$$

In the same way, for N_2 times OOV input, we get

$$y_i = \left[C_{LR} \quad C_{PP} \right], i = 1, 2, \cdots, N_2 \tag{17}$$

We can easily get the expectation of x_i and y_i.

$$\mu_{IV} = \frac{1}{N_1}\sum_i x_i \tag{18}$$

$$\mu_{OOV} = \frac{1}{N_2}\sum_i y_i \tag{19}$$

Now we define within-class scatter matrix S_W :

$$S_W = \sum_i (x_i - \mu_{IV})(x_i - \mu_{IV})^T + \sum_i (y_i - \mu_{OOV})(y_i - \mu_{OOV})^T \tag{20}$$

According to Fisher Rule, we get α and β :

$$[\alpha \quad \beta] = (\mu_{IV} - \mu_{OOV})S_W{}^{-1} \tag{21}$$

Then, appropriate threshold CM_{TH} is chosen, if $CM(W_0, X)$ of (15) exceed the threshold, the output is acceptable. Otherwise, the result is not so believable, and system should reject it or notify the user input again.

4 Evaluation of Confidence Measures [7]

Once the confidence score has been computed, it is compared to a threshold to determine accept or reject it. Since the input speech data might be IV or OOV, two different types of errors can occur. The first is when an OOV input is accepted, which is called false acceptance error(FAR). The second is rejecting an IV input, which is called false rejection error (FRR).

$$FRR = \frac{False \quad rejected \quad word \quad amount}{Total \quad tries} \tag{22}$$

$$FAR = \frac{False \quad accepted \quad word \quad amount}{Total \quad tries} \tag{23}$$

Obviously, there is a trade-off between the two types of errors depending on the threshold. The trade-off is used to form of Detection Error Tradeoff (DET) curves.

Another traditional criterion is the equal-error-rate (EER). The EER can be computed by adjusting the threshold so that FRR and FAR are equal. The lower EER means better robust of the system.

5 Experimental Results

The training corpus is National 863 standard Mandarin Speech Corpus, which contains large amount continuous voice database. It comprises 83 male records and 83 female records. The reading material is people's daily from 1993 to 1994. The signal noise ratio (SNR) of the corpus is about 30dB.

The evaluation corpus comprises names of people, places and stocks, and contains 10 female records and 10 male records. For each person, 600 sentences (isolated-words) are recorded, which involves IV and OOV speech. The corpus is recorded in office environment and 8K Hz sampled.

In order to test the performance of the back-end for different size recognition set, the testing tasks contain a 50 words item and 300 words item.

5.1 In-Vocabulary Recognition Accuracy of the Baseline

In the baseline, we use 27 dimensions of feature vector, which contains 12-dimension Mel-Frequency Cepstral Coefficients (MFCC), 12-dimension ΔMFCC, the normalized energy and its first and second difference.

To get effective performance, we use a set of biphone models. It involves 358 states and each state is fit with the mixture of 3 Gaussian components.

The recognition accuracy is as shown in Table 1. A good accuracy is acquired for IV input for both 50 and 300 words task.

Table 1. Recognition accuracy for IV input

50 Words Task	300 Words Task
99.00%	96.88%

5.2 System Performance of the Confidence Measures and Their Combination

In the back-end, we list 7 secondary hypotheses to work as online garbages. To satisfy the non-dependence, we choose monophone as the acoustic model and unigram as the language model to generate semi-syllable based posterior probability confidence scores. In our language model, we have 27 consonants and 38 vowels in all.

The experimental results will show the performance of the two confidence scores and their combination for 50 words task and 300 words task separately.

System Performance for the 50 Words Task

We design the recognition set including 50 words that consists of 20 person names, 20 place names and 10 stock names. There are 2000 input speech, including 1000 IV input and 1000 OOV input.

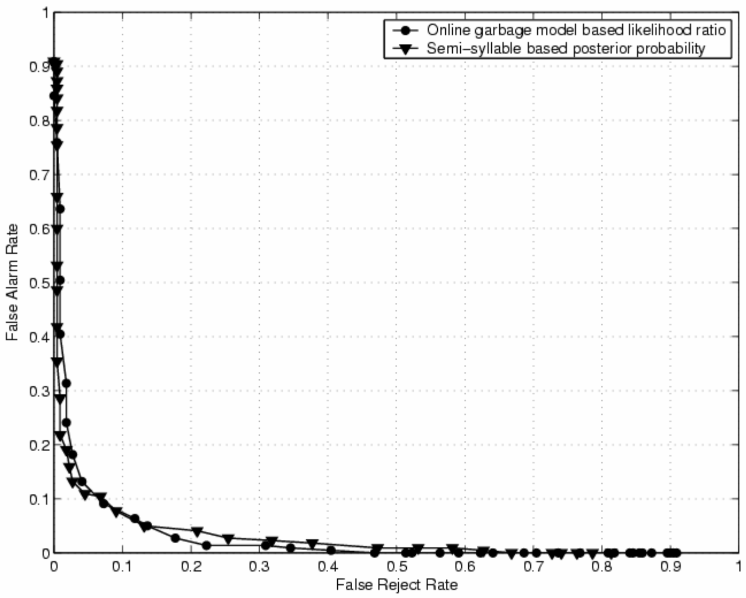

Fig. 1. DET curve of online garbage model based likelihood ratio and semi-syllable based posterior probability for 50 words task

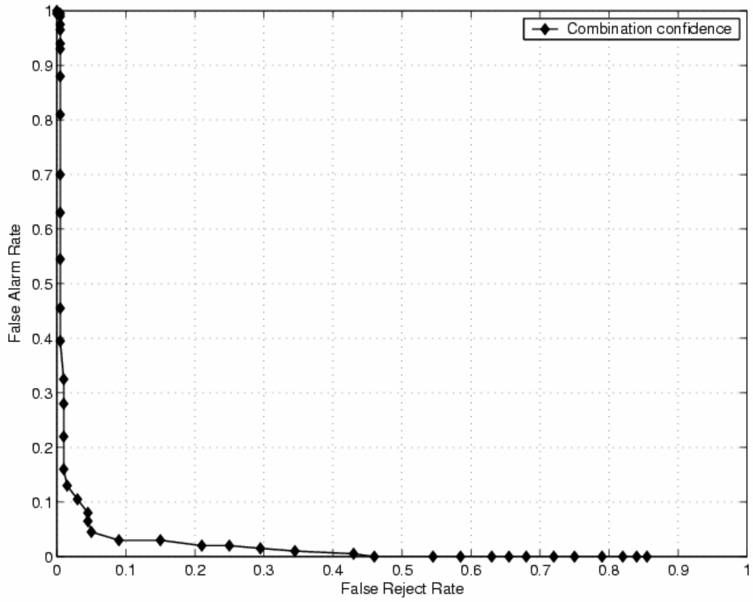

Fig. 2. DET curve of combination confidence for 50 words task

In figure 1, the DET of online garbage model based confidence and semi-syllable based posterior probability confidence for 50 words recognition task are given. From the figure, the EER of the two confidence scores are both of about 9%.

If we combine the two confidence scores, an even better performance is achieved as figure 2 shows. The EER of combination confidence is under 5%. Such back-end processing can be used in some keyword verification system, command recognition system and some other systems of small size recognition task.

System performance for the 300 Words Task

The recognition set includes 300 words: 100 person names, 100 place names and 100 stock names. There are 12000 input speech, including 6000 IV input and 6000 OOV input.

The DET of online garbage model based confidence and semi-syllable based posterior probability confidence for 300 words recognition task are shown in figure 3. The EER of the two confidence scores are of 15% and 12%.

The combination performance is as figure 4 shows and the ERR is of 10%. The performance is acceptable for voice dialing system, stock and airline demanding systems and so on.

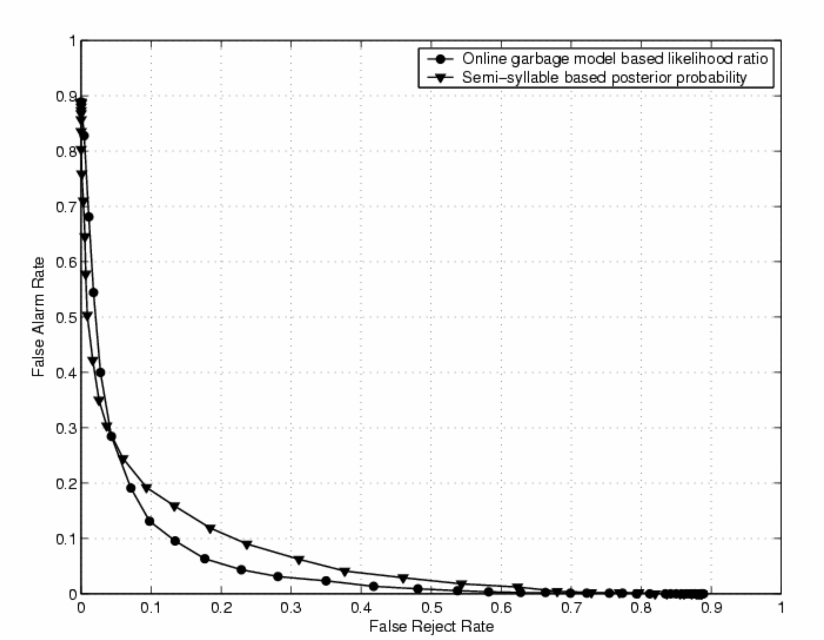

Fig. 3. DET curve of Online garbage model based likelihood ratio and semi-syllable based posterior probability for 300 words task

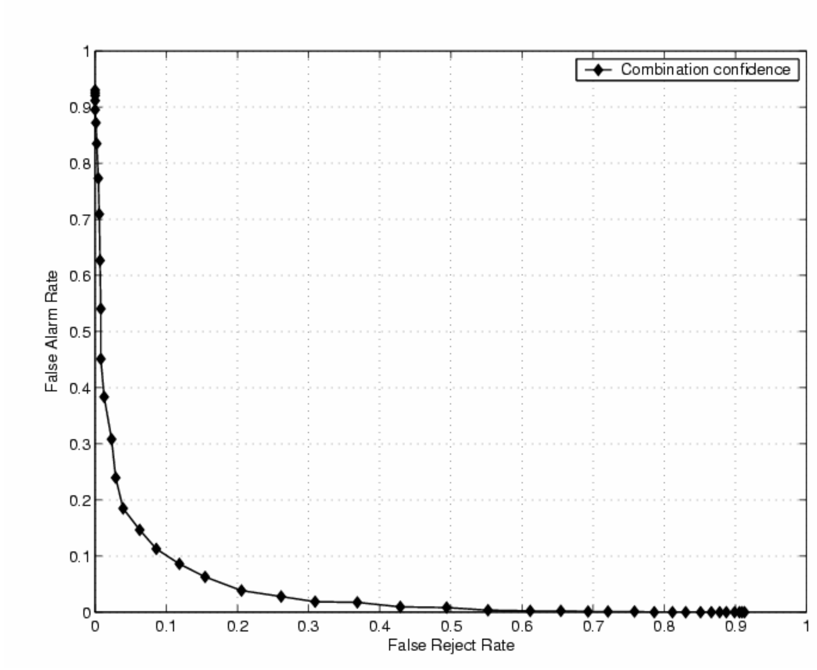

Fig. 4. DET curve of combination confidence for 300 words task

6 Conclusion

This paper studies the back-end of isolated words recognition system. Two types of confidence measures are introduced: online garbage model based likelihood ratio and semi-syllable based posterior probability. Experimental results show that the two confidence measures have an acceptable performance for different size tasks. If we combine the two confidence scores together, an even better improvement has been achieved. For 50 words task and 300 words task, the EER is of 5% and 10%. The back-end can be accepted in practical applications.

Acknowledgment

This project is supported by the National Natural Science Foundation of China (NSFC, 60572083).

References

[1] Zhu Xuan, Chen Yining Liu Jia and Liu Runsheng, A novel efficient decoding algorithm for CDHMM-based speech recognizer on chip. IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 2, April (2003), pp. 293-296.

[2] G. Williams and S. Renals, Confidence measures for hybrid HMM/ANN speech recognition. Proceedings of Eurospeech-97, Vol. 4, September (1997), pp. 1955-1958.

[3] T. Schaaf and T. Kemp, Estimating Confidence Using Word Lattices. IEEE International Conference on Acoustics, Speech, and Signal Processing. Vol. 2, April (1997), pp. 875-878.

[4] M. Weintraub, F. Beaufays, Z. Rivlin, Y. Konig, and A. Stolcke, Neural-network based measures of confidence for word recognition. IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 2, (1997), pp. 887-890.

[5] A. Sankar and S. Wu, Utterance Verification based on statistics of phone-level confidence Scores. IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, (2003), pp. 584-587.

[6] O. Duda, E. hart and G. Stork, Pattern Classification (Second Edition in Chinese). China Machine Press, (2003), pp. 96-99.

[7] F. Wessel, R. Schlter, K. Macherey and H. Ney, Confidence measures for large vocabulary continuous speech recognition. IEEE Transactions on speech and audio processing, Vol. 9, (2001), pp. 288-298.

Yuguo Ding (1980-), male, postgraduate, research interests: speech signal processing and speech recognition.
Jia Liu (1954-), male, professor, research interests: speech recognition/synthesis, chip of speech recognition, multimedia signal telecommunication system.
Runsheng Liu (1933-), male, professor, research interests: digital signal processing, Chinese number signal processing, mixed IC design and speech recognition.